Human vs. Machine: A Study in Choosing the Best Modeling Approach for Your Brand

Given all the available options and industry jargon surrounding data and analytics in real estate forecasting models, there can be confusion around machine learning and artificial intelligence. Machine learning is widely acknowledged as a form of artificial intelligence, but its overuse as a black-box approach in forecasting models can raise questions for real estate planners. Among the available modeling options, how do you know which is right for your brand?

Our team of data scientists conducted a study to test advanced machine learning modeling techniques and determine the effectiveness of each when applied to real estate forecasting for average size brands.

What are machines good for?

Machine learning is a quick and inexpensive substitute for some human abilities. This is especially useful when creating a lot of models or the models need to be rebuilt often. The machine will automatically create a reasonable model for every customer and make a product suggestion to the consumer. This works well as long as the costs of that model making mistakes are minimal.

> *Example: When creating models for a streaming service to recommend a movie, you want to create a model for each customer or update that model every time they watch a new film. If the model keeps recommending "Die Hard" as a "Holiday" selection, even if it's not technically considered a Christmas movie, it doesn't have a big effect on business.*

What can possibly go wrong?

When the costs of making the wrong decisions become higher, it's important to have clarity in understanding how the model functions and make certain it follows patterns, not a few extreme examples. One challenge is that machine learning will pick the most mathematically relevant factors in the model sample without any consideration for the context of these variables.

> *Example: In certain trade areas, a high-end clothing sales model is driven by an average spend on coin laundry. Expensive clothing is correlated to income and coin laundry spend is inversely correlated to income. Consequently the mathematical relationship for coin laundry appears marginally stronger in the model, but the machine has no idea that average income is a considerably more appropriate factor if only slightly less mathematically relevant.*

Sample Size

Machine learning works best and is most efficient when there is a large amount of data to analyze and compare. Large sample sizes can reduce the risks of overfitting and contextual errors. In real estate modeling, an ideal number of locations for a machine-only model would be 10,000 locations or more. However, most retailers and restaurant operators don't have brand portfolios this large—which rules out machine learning alone as the best option. With smaller brands, machine learning can be used to validate analytics, but should not be the only modeling method used.

Humans to the rescue

With machine learning, computers find variables to help predict success or failure and minimize risk, but human input helps determine if these relationships are rooted in real-world effects or simply coincidences found in the sample data. For example, when using modeling for real estate site selection, a human can address the risk of a model adjusting itself in unexpected ways from an event like a natural disaster or the pandemic. Therefore, an experienced statistician who understands how real estate works is recommended to ensure your forecasting model is working properly before you make million dollar decisions with it.

Overfitting is the greatest problem with most machine learning algorithms. The model predicts the training data (or known data) too well. Then when new observations are introduced, most of the predictive ability disappears. You can end up with a model that is only reliable in predicting the past, which is not helpful when trying to determine how a new scenario will perform. You want a model that is good at predicting data you do not have yet.

> *Example: A student memorizes the solution to a problem in a math textbook for an exam without studying how to solve the problem. When the teacher asks to solve a similar but slightly different example, the student only knows the textbook example and cannot generalize to the new problem.*

In the case of overfitting, holdout testing and cross validation are reliable machine tests to identify this problem. With the holdout method the data is divided into subsets, where one set is used to train the model, and the other set as a holdout. We compare the performance of the model on the data that the model has never seen. Cross validation is a machine learning technique to test the models for overfitting when a holdout sample is not available. The sample is divided into multiple subsets where each set is held out one at a time, while the remaining sets are used to train the model. The model is tested on each holdout set.

The best cure for overfitting is a statistician to test each relationship between the variables to make sure the models created are based on patterns in the data and not extreme examples.

Our study of machine learning vs. human modeling

First, we looked at some of the most commonly used modeling algorithms.

Statistical Regression

Statistical regression focuses on the relationship between one dependent variable and a series of other independent variables. It is particularly useful for predicting a new observation by placing it on a linear equation based on existing observations. The linear equation is calculated to minimize the error of the prediction versus the actual. This modeling technique is a combination of machine learning and human supervision.

Nearest Neighbor

Also called a peer comparison or analog method, nearest neighbor compares a new observation to similar peers. In real estate analysis, existing units are identified as the most similar based on difference in key decision variables. If a proposed site in question looks like many other existing locations, you may conclude the new site's performance is likely to be similar. The peer group results are not predictive, but they can be used to validate the statistical regression output.

Decision Trees

A decision tree is a structure that classifies data using a branching method to illustrate possible outcomes of a decision. Each branch within the tree represents testing a specific variable – and each leaf or a node is the outcome of that test. A decision tree model picks the most useful branches to split up data into best predicting final leaves. In real estate, a new observation is predicted by deciding which node of similar units it falls into.

Random Forest

Random forest consists of multiple decision trees that are each slightly different but operate as an ensemble. Each individual tree in the random forest produces a prediction and all the predictions combined become the model's final output.

When comparing the algorithms in the study, our data scientists found the following advantages and disadvantages.

	STATISTICAL REGRESSION	NEAREST NEIGHBOR	DECISION TREES	RANDOM FOREST
Advantages	interpretability, simplicity of final model, statistical testing	highly intuitive, interpretability	intuitive, some interpretability, variables self-selection, deals with nonlinear and outliers well	deals with nonlinear and outliers well, variables self-selection
Disadvantages	stringent data distribution requirements	not built on back checking accuracy	often overfits the data and fails on a holdout, no directionality of factors	limited interpretability, high complexity, can overfit the data and fail on a holdout, no directionality of factors

Methodology and Observations

In the study, our data scientists tested how a human-built, statistical regression model compared to strictly machine-learning built models for a restaurant brand with over 900 locations. We then compared the model results to the brand's actual performance. This exercise produced the following observations:

- The machine learning models rarely beat human-built regression models on the cross validation or holdout accuracy.
- Given more variables, only the random forest machine learning technique beat the human by a small margin.
- Almost identical groups of variables are identified as key success drivers despite the modeling technique.
- All models give a relative importance of the variables, but with machine learning, the direction of the variable's effect is lost. Any given variable can be used multiple times, and sometimes it can have positive effect on sales and in other parts of the tree, a negative effect.

Conclusion

In real estate forecasting, more advanced models can provide another useful decision point when trying to predict future performance, but they do not outperform the predictive models built by humans. There's a lot at stake in choosing a new location for your brand. A strictly machine learning approach can provide black-box predictions, but a customized solution using human-built predictive models provides clarity into unique factors driving success to help confidently select locations that will thrive.

At CBRE Forum Analytics, we do more than deposit data into standard algorithms. Our models are curated and validated for each client by using the appropriate algorithms for the data we have. To learn more, schedule a demo today.